Kapitola 9. A

Regresná analýza

V kapitole 8. A sme analyzovali dvojrozmerný štatistický súbor a počítali korelačný koeficient ako kvantitatívnu mieru vzájomnej závislosti. Ak bol korelačný koeficient ρ (výberový korelačný koeficient r_{xy}) blízky jednej alebo mínus jednej bola medzi premennými lineárna závislosť. Našou snahou je popísať túto závislosť nejakou funkciou, najčastejšie polynómom prvého alebo druhého stupňa t. j. pre dvojrozmernú náhodnú premennú (X, Y) nájsť funkčnú závislosť

$$Y = f(x, a_0, ..., a_k)$$

a odhadnúť neznáme parametre a_0, ..., a_k. Metóda, ktorú popíšeme sa nazýva ***metóda najmenších štvorcov***. Odhady $\hat{a}_0, ..., \hat{a}_k$ neznámych parametrov a_0, ..., a_k volíme tak, aby sa minimalizovali chyby, ktoré predstavujú rozdiely medzi teoretickými a skutočnými (nameranými) hodnotami premennej Y. Keďže rozdiely môžu nadobúdať ako kladné, tak aj záporné hodnoty, umocňujú sa druhú a počítajú sa ich súčty. Minimalizujeme vlastne T-súčet kvadratických odchýlok (štvorcov) teoretických a skutočných hodnôt premennej Y.

$$T(\hat{a}_0, ..., \hat{a}_k) = \sum_{i=1}^{n} (y_i - f(x_i, \hat{a}_0, ..., \hat{a}_k))^2$$

Funkcia f sa nazýva ***regresná funkcia***. Ak má konečné prvé parciálne derivácie podľa všetkých parametrov a_0, ..., a_k, potom odhady $\hat{a}_0, ..., \hat{a}_k$ získame riešením (k +1) rovníc o (k +1) neznámych:

$$\frac{\partial T}{\partial a_i} = 0 \quad pre \quad i = 0, ..., k$$

Niektoré regresné krivky sa v praxi veľmi často používajú a preto nájdeme presný tvar odhadov $\hat{a}_0, ..., \hat{a}_k$ (resp. zostavíme systém rovníc z ktorého ich možno nájsť).

- **<u>Lineárna regresia</u>**

Regresná funkcia má tvar

$$y_i = a_0 + a_1 * x_i \qquad i = 1, ..., n$$

Riešením príslušnej sústavy dvoch rovníc o dvoch neznámych sú odhady v tvare

$$a_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

- **Kvadratická regresia**

Regresná funkcia má tvar

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 \qquad i = 1,..., n$$

Odhady príslušných parametrov získame riešením troch rovníc o troch neznámych, ktoré sú v tvare

$$\sum_{i=1}^{n} y_i = a_2 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i + n a_0$$

$$\sum_{i=1}^{n} x_i y_i = a_2 \sum_{i=1}^{n} x_i^3 + a_1 \sum_{i=1}^{n} x_i^2 + a_0 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i^2 y_i = a_2 \sum_{i=1}^{n} x_i^4 + a_1 \sum_{i=1}^{n} x_i^3 + a_0 \sum_{i=1}^{n} x_i^2$$

Vzorový príklad 1

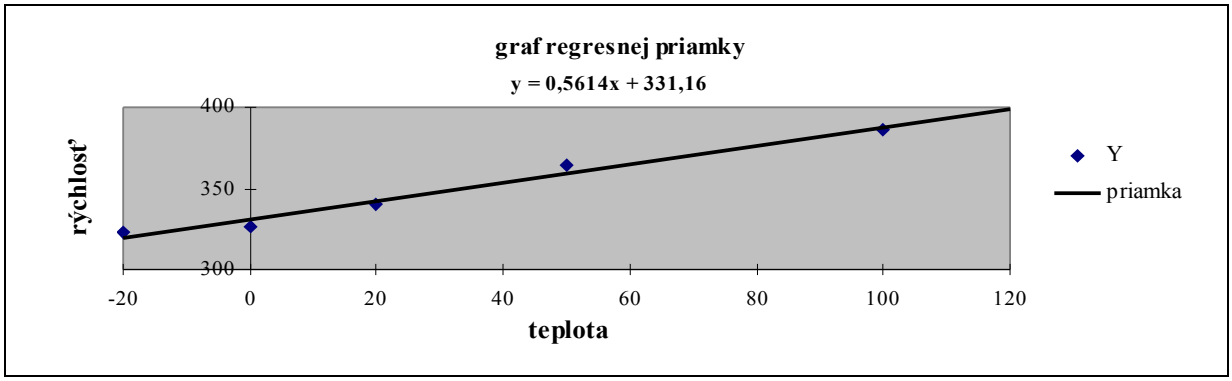
Merala sa rýchlosť zvuku y_i pri rôznych teplotách vzduchu x_i, výsledky sú zaznamenané v tabuľke.

teplota (^{0}C)	-20	0	20	50	100
rýchlosť (m/s)	323	327	340	364	386

Vypočítajte korelačný koeficient, preložte regresnú priamku.

Riešenie

Postupne vypočítame a dosadíme do vzorcov pre odhad parametrov regresnej priamky. Výberový korelačný koeficient sa rovná r_{xy} = 0,989 a odhady parametrov regresnej priamky postupne $\hat{a}_0$ =331,159, $\hat{a}_1$=0,561. Graf regresnej priamky pozri na obrázku 9.1 (výstup z Excelu).



Obr. 9.1

Vzorový príklad 2

V desiatich rôznych vzdialenostiach x_i sa merala veľkosť priehybu zaťaženia dosky y_i . Výsledky sú uvedené v tabuľke.

x_i (dm)	1	2	3	4	5	6	7	8	9	10
y_i (mm)	2,1299	2,1532	2,1611	2,1516	2,1282	2,0807	2,0266	1,9594	1,8759	1,7723

Preložte regresnú krivku druhého stupňa (kvadratická regresia).

Riešenie

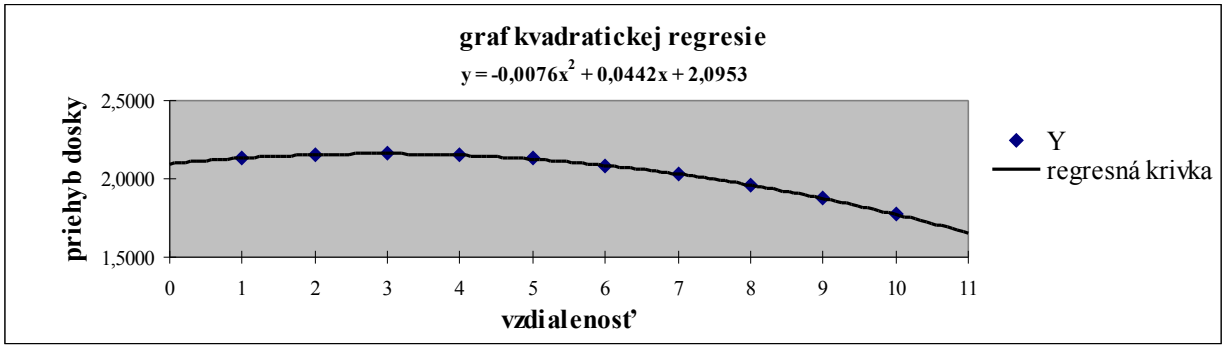
Vypočítame príslušné sumy a zostavíme systém rovníc

$$20{,}4389 = 385a_2 + 55a_1 + 10a_0$$
$$109{,}1187 = 3025a_2 + 385a_1 + 55a_0$$
$$746{,}6113 = 25333a_2 + 3025a_1 + 385a_0$$

Riešením tejto sústavy rovníc sú odhady parametrov $\hat{a}_0$ =2,09530, $\hat{a}_1$ =0,04420, $\hat{a}_2$ =-0,00765. Hľadanou závislosťou veľkosti priehybu od vzdialenosti je parabola v tvare

$$y = 2{,}09530 + 0{,}04420x - 0{,}00765x^2$$

Graf regresnej krivky pozri na obrázku 9.2 (výstup z Excelu).



Obr. 9.2

Transformácie na lineárnu regresiu

Mnohé funkcie, ktoré nie sú lineárne, možno jednoduchou transformáciou na lineárne previesť a tak použiť metódu najmenších štvorcov. Príklady niektorých transformácií sú uvedené v nasledujúcich riadkoch

- $y = ae^{bx}$, transformácia $\ln(y) = \ln(ae^{bx}) \Rightarrow z = A + Bu$, kde $z = \ln y \;\; u = x, \;\; A = \ln a$, $B = b,$ pôvodné parametre sú $a = e^A$, $B = b$
- $y = ax^b$, transformácia $\ln(y) = \ln(ax^b) \Rightarrow z = A + Bu$, kde $z = \ln y, \;\; u = \ln x, \;\; A = \ln a,$ $B = b,$ pôvodné parametre sú $a = e^A$, $B = b$
- $y = ab^x, b > 0, b \neq 1$, transformácia $\ln(y) = \ln(ab^x), \Rightarrow z = A + Bu$, kde $z = \ln y, \;\; u = x,$ $A = \ln a$, $B = \ln b,$ pôvodné parameter sú $a = e^A$, $b = e^B$
- $y = a + \frac{b}{x},$ transformácia $x = \frac{1}{u}, \Rightarrow z = A + Bu$, kde $z = y, \;\; u = \frac{1}{x}, \;\; A = a, \; B = b$

- $y = \frac{1}{a+bx}$, transformácia $y = \frac{1}{z}, \Rightarrow z = A + Bu$, kde $z = \frac{1}{y}$, $u = x$, $A = a$, $B = b$

Vzorový príklad 3

Barometrický tlak p (meraný v Pa) závisí exponenciálne od nadmorskej výšky h (meranej v m) t.j.

$$p = ae^{bh}$$

namerali sme 6 hodnôt barometrického tlaku v rôznych nadmorských výškach. Výsledky sú uvedené v tabuľke:

H	0	270	840	1452	2116	3203
P	100000	96974	90263	83553	76842	66842

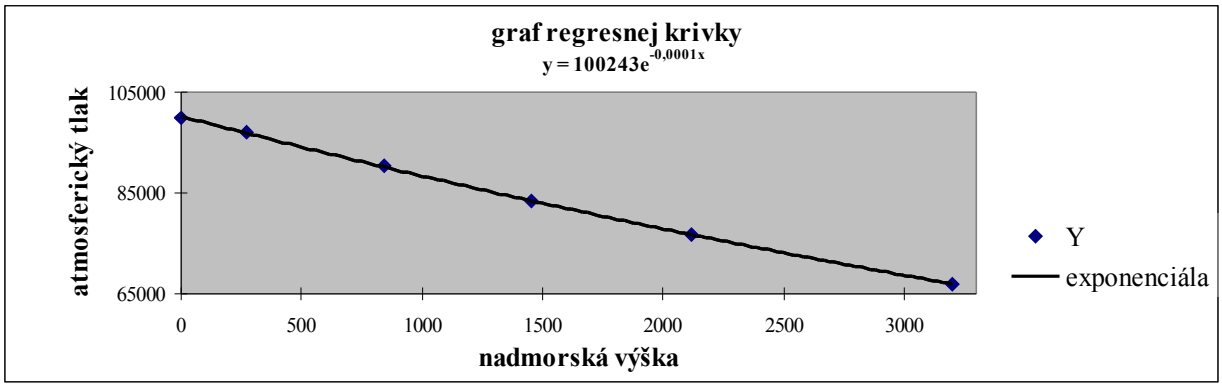
Vhodnou transformáciou linearizujte, metódou najmenších štvorcov odhadnite parametre a, b.

Riešenie

Požijeme vzorce pre transformáciu $y = ae^{bx}$, t.j. $\ln(y) = \ln(ae^{bx}) \Rightarrow z = A + Bu$ kde $z = \ln y$ $u = x$, $A = \ln a$, $B = b$, pôvodné parametre sú $a = e^A$, $B = b$, dosadíme príslušné hodnoty do vzorcov pre lineárnu regresnú priamku, A = 11,515, B = -0,000126 ⇒ a = $e^{11,515}$ = 100243,51 , B = b = -0,000126. Vzťah závislosti tlaku od nadmorskej výšky udáva funkcia

$$p = 100243{,}51e^{-0{,}000126h}$$

Graf regresnej krivky pozri na obrázku 9.3 (výstup z Excelu).



Obr. 9. 3

Poznámka 1

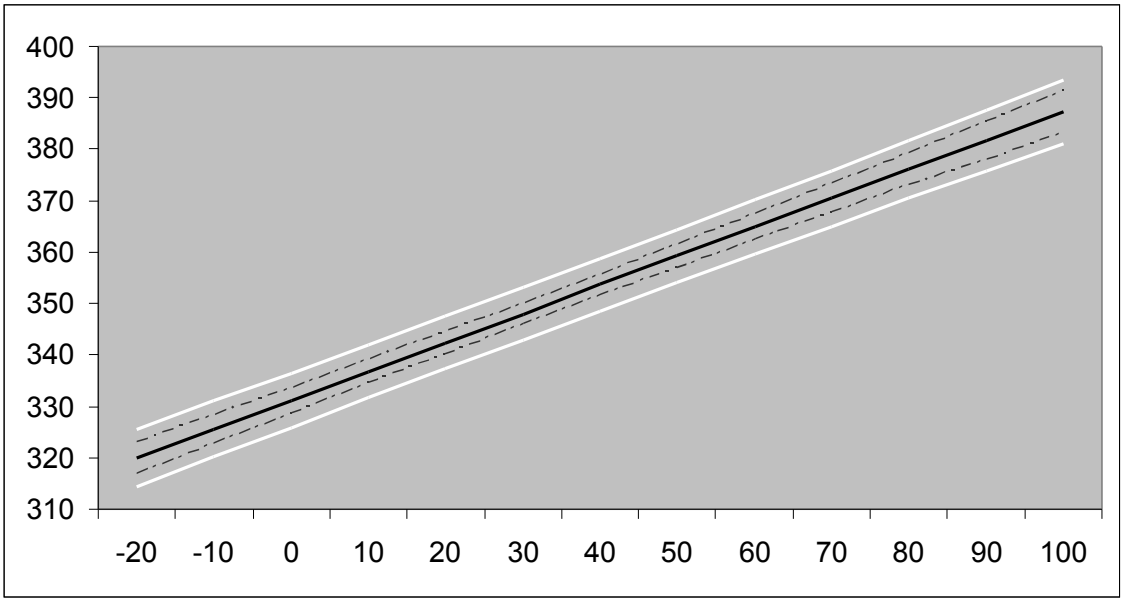
Pre regresnú priamku $y = a_0 + a_1 x$ sa zvyknú počítať a vykresľovať dva pásy, ktoré sa nazývajú ***interval spoľahlivost pre regresnú priamku*** a ***interval spoľahlivosti okolo regresnej priamky***. (1-α) 100% interval spoľahlivosti pre regresnú priamku je interval spoľahlivosti v tvare:

$$\left(a_0 + a_1 x - t_{n-2}(1-\frac{\alpha}{2}) s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}; a_0 + a_1 x + t_{n-2}(1-\frac{\alpha}{2}) s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \right)$$

pričom $t_{n-2}(1-\frac{\alpha}{2})$ je príslušný kvantil t rozdelenia a $s^2 = \frac{\sum_{i=1}^{n}(y_i-(a_0+a_1x_i))^2}{n-2}$. Ak postupne rátame hodnoty ľavého a pravého koncového bodu intervalu v meniacich sa hodnotách x, pospájame zvlášť ľavé a pravé koncové body, dostaneme pás spoľahlivosti pre regresnú priamku. Interval spoľahlivosti okolo regresnej priamky vznikne tak, že pre konkrétnu hodnotu x vypočítame príslušný interval spoľahlivosti tvare:

$$\left(a_0+a_1x-t_{n-2}(1-\frac{\alpha}{2})s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum x_i^2-n\bar{x}^2}};\ a_0+a_1x+t_{n-2}(1-\frac{\alpha}{2})s\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum x_i^2-n\bar{x}^2}}\right)$$

kde opäť $t_{n-2}(1-\frac{\alpha}{2})$ je príslušný kvantil t rozdelenia a $s^2 = \frac{\sum_{i=1}^{n}(y_i-(a_0+a_1x_i))^2}{n-2}$. Napriek tomu, že rátame interval spoľahlivosti pre jediné x, zvyknú sa opäť pospájať ľavé a pravé koncové body týchto intervalov pre meniace sa hodnoty x a tak vznikne pás, korý voláme pásom spoľahlivosti okolo regresnej priamky. Na záver uvedieme graf regresnej priamky y=331,159+0,561x (vzorový príklad 1) spolu s príslušnými pásmi spoľahlivosti pre regresnú priamku (šedá bodkočiarkovaná čiara) a okolo regresnej priamky (biela plná čiara).



Obr. 9.4

Nové pojmy a definície kapitoly 8. A

- *regresná funkcia*
- *metóda najmenších štvorcov*
- *normálne rovnice*
- *parametre lineárnej regresie, odhady parametrov*
- *lineárna a kvadratická regresia*

- *transformácia na lineárnu regresiu*
- *interval spoľahlivosti pre regresnú priamku*
- *interval spoľahlivosti okolo regresnej priamky*

Cvičenia

Príklad 1
Pri hodnotení skúšok na únavu materiálu možno popísať závislosť počtu kmitov do lomu y na napätí x vhodnou regresnou priamkou. Tabuľka udáva hodnoty napätia x (v MPa) a y (počet kmitov).

X	560	580	600	630	650	700
Y	2845	2597	1227	1200	728	600

Vypočítajte korelačný koeficient. Nájdite odhady parametrov regresnej priamky.

Príklad 2
Firma vyrábajúca stavebné náradie vybrala 6 dealerov za účelom zistenia ročných zásob. Objem zásob v minulom roku označme x a v tomto roku y. Výsledky sú zaznamenané v tabuľke

X	70	260	150	100	20	60
Y	6	320	230	120	50	60

Vypočítajte korelačný koeficient, preložte vhodnú regresnú priamku, odhadnite stav tohtoročných zásob v podniku, kde v predchádzajúcom roku bol objem zásob x = 200.

Príklad 3
Merala sa hustota vody y (v kg/dm^3) v závislosti na teplote x (v ^0C). Výsledky meraní sú uvedené v tabuľke

X	10	20	30	40	50	60
Y	1	0,997	0,996	0,993	,987	0,983

Vypočítajte korelačný koeficient, preložte vhodnú regresnú krivku (parabola). Akú hustotu má voda pri bode varu a bode mrazu?

Príklad 4
Sledujeme priehyb y ($mm*10^{-2}$) plastickej hmoty v závislosti na tlaku x (kp/cm^2). Získané údaje sú uvedené v tabuľke

X	4	6	8	10	12	14	16
Y	18	35	48	61	80	93	94

Vypočítajte korelačný koeficient, preložte vhodnú regresnú krivku, určite pri akom tlaku sa plast zlomí, ak hraničná hodnota bodu zlomu je priehyb 100.

Príklad 5
Merala sa závislosť koeficientu viskozity y pri prúdení kvapaliny potrubím na Reynoldsovom čísle Re. Získané výsledky sú uvedené v tabuľke

Re	10000	20000	50000	100000	150000	200000
Y	0,032	0,026	0,021	0,018	0,016	0,015

Zo skúseností vieme, že funkčná závislosť y od Re je daná vzťahom $y=a*Re^b$. Nájdite odhady a, b. Vypočítajte korelačný koeficient.

Riešenie

Príklad 1
r_{xy}=-0,89, y=a +bx, kde a=11895,42 a b=-16,713

Príklad 2
r_{xy}=0,937, y=a+bx, kde a=-14,84 a b=1,325 y(200)=250,324

Príklad 3
r_{xy}=-0,974, $y=ax^2+bx+c$, kde a=-0,000004 a b=-0,00002, c=1,003, y(0)=1,003, y(100)=0,9979

Príklad 4
r_{xy}=0,988, $y=ax^2+bx+c$, kde a=-0,202 a b=10,762, c=22,857, $100=-0,202x^2+10,762x+22,857$

Príklad 5
r_{xy}=-0,891, $y=ax^b$, x=Re, a=0,3108, b=-24,85